An Examination of the Architecture and System-level Tradeoffs of Employing Steep Slope Devices in 3D CMPs

Karthik Swaminathan Huichu Liu Jack Sampson Vijaykrishnan Narayanan

The Pennsylvania State University
{kvs120,hxl249,sampson,vijay}@cse.psu.edu

Abstract

For any given application, there is an optimal throughput point in the space of per-processor performance and the number of such processors given to that application. However, due to thermal, yield, and other constraints, not all of these optimal points can plausibly be constructed with a given technology. In this paper, we look at how emerging steep slope devices, 3D circuit integration, and trends in process technology scaling will combine to shift the boundaries of both attainable performance, and the optimal set of technologies to employ to achieve it. We propose a heterogeneous-technology 3D architecture capable of operating efficiently at an expanded number of points in this larger design space and devise a heterogeneity and thermal aware scheduling algorithm to exploit its potential. Our heterogeneous mapping techniques are capable of producing speedups ranging from 17% for a high end server workloads running at around 90°C to over 160% for embedded systems running below 60°C.

1. Introduction

Modern processor designers have sufficient transistor resources to include an increasing number of increasingly faster cores in a processor design, but lack the power budgets to scale aggressively in both per-core performance and number of cores simultaneously [40, 12] – although peak performance may still scale [30], sustainable performance is increasingly constrained. Notably. for any given application, there are one or more optimal throughput points in the space of per-processor performance and the number of such processors given to that application. However, due to thermal, yield, and other constraints, not all of these optimal points correspond to currently realizable systems. Similarly, for metrics outside of performance, such as average power, total energy, or silicon area, these points may be quite diverse.

While Moore's law continues to advance for the time being, promising access to progressively more parallel systems, limitations on the balance between leakage and dynamic power, wires and logic, and complexity and power density fundamentally couple the ability to realize superior design points in current and future process generations with the energy efficiency of the processors employed. With the end of Dennard scaling [8], there is increased interest in techniques [10] and technologies [27, 29, 17] that promise fundamentally more energy efficient computation. However, many of these CMOS alternatives offer their benefits in energy/instruction at the cost of greatly reduced performance. NEMS relays, for example, have effectively zero leakage power, but can only operate at frequencies orders of magnitude slower than CMOS gates.

In this paper, we focus on the impact that emerging devices and techniques will have on shifting which regions of the design space correspond to systems that are both plausible for mass deployment (i.e. they can be produced with meaningful yield and operate at commonly accepted peak temperatures) and preferable (e.g. among designs that can achieve the same performance in multiple implementations, we would prefer the more efficient or the cheaper of any two designs). In particular, we focus on the rapidly maturing technique of 3D integration [13, 5] and the potential benefits offered by designs built with *Interband Heterojunction Tunnel Field Effect Transistors* (*TFETs*) [29, 34, 33].

Both 3D integration and TFET designs offer the potential to extend the maximum number of aggressive cores possible within a viable yield and thermal budget. Yield decreases superlinearly with increases in area [5], and communication costs among cores scale poorly in planar designs [41]. Thus, 3D integration offers a very direct means to achieve meaningfully higher core counts in tightly integrated systems. However, moving to a 3D design aggravates thermal limitations by placing both additional heat sources and insulators between the cooling system and lower layers in the stack. On the other hand, TFETs and other steep-slope devices offer fundamental reductions in leakage currents and switching energy at the cost of a more limited upper range of operating frequencies. There is a natural synergy between 3D integration and TFETs in that reducing the thermal density on each layer by substituting TFET designs for CMOS will allow more layers within the thermal budget, allowing 3D TFET based designs to scale to sufficient parallelism to overcome limitations in the serial performance of TFET based processors. However, while deploying TFET based designs in a 3D architecture is conceptually appealing, many questions regarding how best to design such a multiprocessor (e.g. microarchitecture selection, performance targeting, scaling) have not been definitively answered.

The contributions of this paper are as follows:

- We conduct an extensive evaluation of the performance and energy tradeoffs among choices in device technologies, 3D integration, microarchitecture, and scheduling under constraints imposed by realistic yield models, thermal bounds

978-1-4799-4394-4/14/$31.00

and exploitable application parallelism.

- We show that, with 3D integration, steep-slope based devices are already plausible candidates for achieving peak performance for highly parallel applications.
- We show how, with further technology scaling, the range of applications for which steep-slope devices are appropriate grows, while the portion of the design space where CMOS is optimal shrinks to a point where only a small number of CMOS cores may be desirable.
- We present an intelligent scheduling approach for hybrid CMOS-TFET systems that allows less parallel applications to still achieve a significant fraction of their peak performance on a primarily TFET-based system.

The remainder of the paper proceeds as follows. Section 2 motivates the opportunities for new technologies to expand the realm of viable latency and parallelism tradeoffs Section 3 provides background information on the properties of TFETs and the modeling of TFET-based designs. Section 4 describes our approach to exploring the design space, Section 5.3 details our methodology, and Section 6 presents the results of our investigations. Section 7 reviews related work and Section 8 concludes.

2. Motivation

Broadly, performance improvements in general-purpose cores can be realized in two dimensions - by reducing single threaded latencies via increasing the frequency or by exploiting the inherent parallelism (TLP and ILP) of the application by increasing the number of application cores or architectural complexity (issue width, pipeline stages etc.). In theory, the increase in frequency can continue until fundamental physical properties of the transistors and wires allow it. Similarly, the increases from core counts are only restricted by the scalability of the application. However, in reality, there are several other constraints that crop up far earlier. Every processor is limited by the total power consumed, namely its power budget, which restricts the attainable processor configurations. This problem can be mitigated to an extent by various approaches [38] for exploiting *Dark Silicon* i.e by spatially or temporally reallocating power budgets such that either subsets of (possibly specialized) cores can operate at peak frequency or all cores can operate at peak frequency a subset of times at the expense of darkening/dimming other cores/times.

In addition to power, there are two other key considerations for understanding which processor configurations are practical. Namely, yield constraints may restrict the manufacturability of processors with high core counts [11] and *thermal limitations* due to power density may come into play even for processors staying within their aggregate power budget. Below, we examine these two constraints in more detail and then discuss how emerging devices help us expand our design space in light of these constraints.

2.1. Thermal constraints on processor execution

Power budgeting has become an important consideration in the design and operation of processors. This power constrained operation can extend across a wide range of application domains, ranging from the mobile and embedded space to the high-end server space. However, constraining total power does not enforce adherence to the inherent thermal limitations of processor components. The component temperature depends, not on power, but on power density. Most processor components are rated to operate within a fixed range of temperatures and exceeding this temperature range can have an adverse impact on their lifetime and reliability. The *Thermal Design Power* or TDP is an indication of the peak power level that the processor can achieve without causing the thermal limit to be crossed.

2.2. Yield constraints on processor design

To exploit application parallelism, increasing per-chip core count can be done without significantly aggravating power density. However, increasing the die size can adversely affect the overall processor yield, since the yield is inversely proportional to the chip area. Folding cores onto multiple layers (e.g. 3D stacked chips) can reduce the area footprint. While this has ramifications both in increasing the processor yield as well as improving on-chip bandwidth and latency due to reduced interconnect length, there is a price to pay, in terms of bonding yield, for increasing the number of layers, and this limits returns on increasingly stacked chips. Further, increasing layer counts exacerbates thermal limitations, since the inner layers lack an efficient means for heat dissipation.

Figures 1a) and b) show the extent of frequency and core scaling for two applications, *barnes*, which scales well, and *ocean.cont*, which scales poorly. The regions shaded black correspond to the points at which the scaling model "collapses", i.e thermal and yield considerations restrict the design space. While both applications are affected by the frequency limitation, only *barnes* is adversely affected by the constraint on the number of cores.

2.3. Opportunities with TFET processors

As Section 3 describes in detail, TFET cores can provide a more energy efficient alternative to conventional CMOS processors, especially at near-threshold and sub-threshold voltage – at sufficiently low voltages, the steep slope of TFETs makes them inherently more efficient transistors independent of process tuning that can be done to customize CMOS [21]. Substituting TFET cores for CMOS cores lessens the thermal consequences of 3D stacking. Consequently, stacked TFET cores extend the range of viable designs in the core count/frequency space. Similarly, operating CMOS cores at increased supply voltage (V_{dd}) enables high frequency operation.

There are several avenues to explore in order to trade-off the lower temperature operation of TFETs for increased performance. In this paper we focus on the advantages of extend-

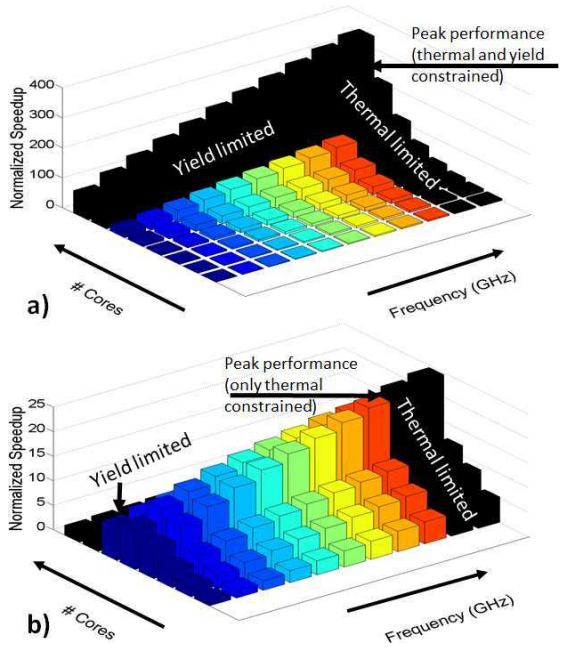

Figure 1: a) and b) Demonstration of yield and thermal limits on performance scaling in the frequency (X) domain and the parallelism (Y) domain for a well scaling application (*barnes*) and poorly scaling application (*ocean.ncont*)

ing device-level heterogeneity to 3D stacked processors, thus aiming to increase the design-space boundary illustrated in Figure 1. The two main roadblocks encountered in this effort are the decrease in yield due to bonding and TSV losses, and the steady increase in power density as layers are added, leading to large temperature increases among the internal layers.

3. Modeling TFET-based processors

In this section, we provide a basic introduction to the benefits and limitations of the steep-slope devices we consider, and describe how we extrapolate from these effects to processor-level models.

3.1. Basic background on Steep-slope devices

Steep slope devices have been proposed as an alternative to counter the 60 mv/decade subthreshold limitation that restricts the scaling of conventional CMOS transistors. This property of CMOS causes an exponential increase in leakage current as supply voltage (V_{dd}) scales to near- and subthreshold values and threshold voltage (V_T) remains roughly constant. In contrast, steep-slope devices do not suffer from this sub-threshold limitation on account of their charge transport mechanism which involves tunneling through the intrinsic region. Hence, at near-threshold and subthreshold voltages, these steep slope devices have the potential to outperform CMOS devices to several orders of magnitude.

Heterojunction Tunnel FETs (TFETs) are one of the most promising step slope devices, both in terms of subthreshold operation and high speed switching operations, and TFET devices have been shown to scale well into future process nodes [26]. There have been experimental demonstrations for III-V n-FET and SiGE p-FET co-integration in [7] and for vertical III-V TFET integration on Si substrate in [31] and [39], which makes heterogeneous integration of TFET and CMOS devices possible on the same die. TFETs do however, suffer in some aspects when compared to CMOS technology. As the supply voltage is increased, the inherent limitation in the TFET charge-carrying mechanism causes the current to saturate above a certain operating voltage. Due to the saturation of the tunneling current, the switching delay remains constant beyond a certain supply voltage. At a processor level, this translates to a limited range of operating frequencies for TFET transistors.

It is possible to tune TFET device characteristics to an extent to improve frequency and power responses through altering channel length [25]. By using a low static power (LSTP) TFET with an increased channel length, our TCAD [1] simulations show that we will able to realize the drive current required while drastically reducing overall power consumption by 2× over existing device models [21].

3.2. Extrapolation to processor model

We use McPAT [24] to obtain the total processor power for an equivalent FinFET transistor based technology. We obtain the corresponding TFET core power by scaling this obtained value with transistor level delay and power parameters.

Such a simple scaling, does not, however, take into account the interconnect delay and power in the processor. Assuming the entire processor to consist solely of transistors subject to CMOS-TFET scaling would result in substantial inaccuracies in the model. Figure 2 a) and b) shows the proportion of wire power and delay to the total core power and critical path delay respectively, for both CMOS and TFET processors of different issue width configurations. In the lower power TFET processors, the increase in wire power with architectural complexity is non-uniform, even though wire lengths increase monotonically with issue width. This is because the leakage power increases non-linearly with respect to the wire power, modifying the relative ratios of dynamic, leakage and wire power. Since these wire parameters are relatively invariant to core frequency, the proportion of wire power is highest at low frequencies (minimum logic power) and the proportion of wire delay is highest at high frequency (minimum logic delay) points. Hence we plot Figure 2 a) and b) at these corresponding extreme points (500 MHz and 2 GHz respectively for a CMOS processor and 500 MHz and 1.5 GHz for a TFET processor). We observe that the contribution of wire delay to the total critical path delay can be significant (up to 30% for CMOS and TFET processor). In a similar manner, the wire power also constitutes nearly 30% of the total processor power in CMOS and nearly 50% of the total power in TFET processors.

While incorporating wire models into our device-to-processor abstraction, we assumed a direct device substitu-

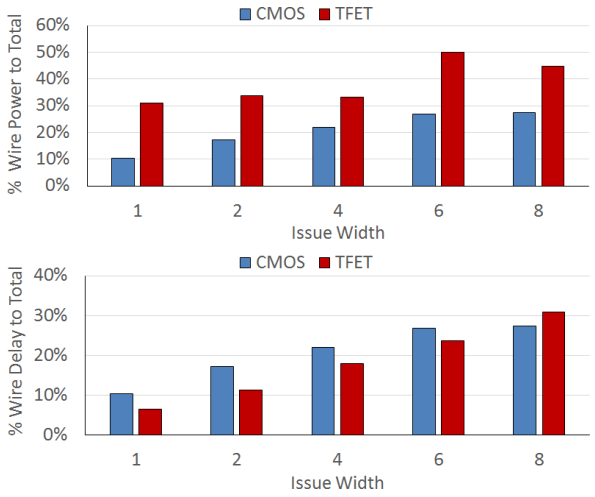

Figure 2: a) and b) Fraction of wire power and wire delay to total core power and delay respectively.

tion from Si FinFET to TFET technology at 22 nm. As a result, we scaled all transistor delays and power for logic and memory components. For memory components, TFET designs have additional structural differences due to their unidirectional conduction. The wire models largely remain the same across technologies, except for the buffer and repeater logic in large wires, which we rescaled according to relative drive strengths.

Figure 3 shows the power-frequency tradeoffs for a 2-issue Atom-like core and a 4-issue Ivybridge-like core, when both are realized using CMOS and TFET technology. We define "crossover frequency" as the operating frequency (and associated voltage) where the TFET and CMOS based designs provide equivalent energy efficiency. The crossover frequency (f_c) point is lower for the Ivybridge core compared to the Atom core. This is because the timing constraints for the complex core (in McPAT) are more severe and the slack allowed to the slower TFET transistor is relatively lower.

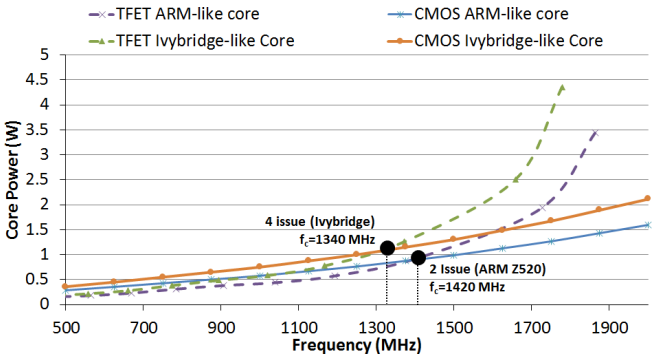

Figure 3: Variation of CMOS and TFET core power with frequency for a simple (Atom-like) core and a complex (Ivybridge-like) core. The crossover point is seen to shift to the left as core complexity increases.

4. Design Space Analysis

Table 1 describes the various dimensions we evaluate in the 3D-device-heterogeneous architecture space.

CMOS Technology	22nm Si FinFET
CMOS Frequency Range	0.5 GHz - 3 GHz
TFET Frequency Range	0.5 GHz - 1.75 GHz
TFET Technology	22nm HTFET
Number of layers	1 - 8
Total Number of Cores	1 - 128
Number of utilizable cores	1-64
Thermal limit	360K (air cooled)
Ambient temperature	300K

Table 1: Configuration of the evaluation platform.

In this section, we analyze the effect of each of these dimensions on thermally constrained performance.

4.1. Technology variation

Our base device models are based on 20 nm technology node simulations calibrated with fabricated devices [33]. In addition, we also use TFET device scaling [26] to model technologies at the 10 nm node (where we expect TFETs to become commercially available).

Figure 4 shows the scaling of the critical path delay when extrapolated to future technology nodes. Comparisons are made between the ITRS 2012 [19] roadmap projections for Si FinFET and simulation results for HTFET. The HTFET device models for 14 nm and 10 nm technology nodes are generated from simulations using the TCAD Sentaurus device modeling tool [1]. The supply voltage V_{cc} corresponding to each technology node is $V_{cc} = 0.72V$ (22 nm node), 0.67 V(14 nm node), 0.55 V (10 nm node) for Si FinFET technology; and $V_{cc} = 0.4V$ (22 nm node), 0.35 V (14 nm node) and 0.3 V (10 nm node) for HTFET technology. In a similar manner, Figure 5 shows the scaling of the total core power for each of the above technology nodes.

From our models, we observe that one of the major limitations with TFET processors at the 22 nm node is their relatively low peak performance. The saturating nature of TFET tunneling current forces the peak frequency to be restricted to around 1.5-1.6 GHz, which is far below what CMOS processors are capable of attaining. However the minimum switching delay of the device reduces with subsequent generations, enabling TFET processors to operate at much higher frequency. Although there is a proportional decrease in FinFET switching as well, the non-scaling of wire-delays causes the frequency gap between CMOS and TFET processors to shrink with every generation. This is because the gap in the critical path delay between CMOS and TFET goes on decreasing with technology and by the 10nm node, TFET cores can attain 95% of the peak performance of CMOS, as compared to 60% for the current (22nm) technology node. Further, TFETs become more and more power efficient w.r.t CMOS with each subsequent generation. Thus the range of applica-

tions where TFETs can act as a viable replacement goes on expanding as transistors continue to scale.

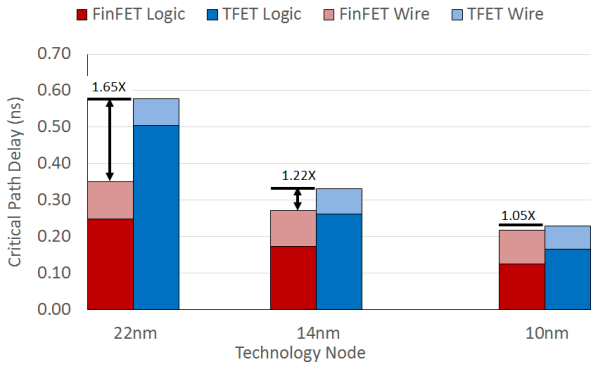

Figure 4: Variation of total critical path delay (logic + wire delay) in FinFET and TFET processors at 22, 14 and 10 nm technology nodes. Scaling is demonstrated both for the ITRS roadmap projections and TCAD simulations for FinFET and TFET.

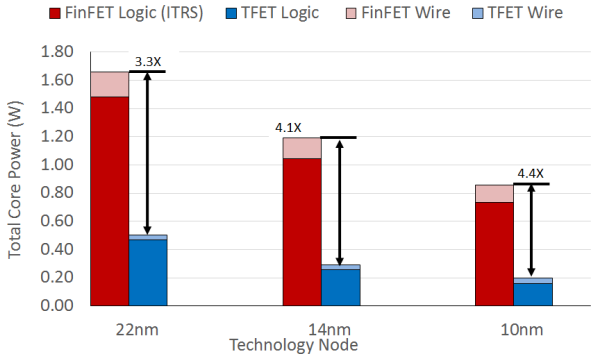

Figure 5: Variation of total core power (including logic and wire power) in a CMOS and TFET processor at 22 nm, 14 nm and 10 nm technology nodes.

4.2. Yield aware stacking of processors

As explained in Section 2, processor yield reduces super-linearly with increase in die area. As shown in equation 21 in [5], the yield varies with area as a Gamma function. While reducing the area footprint by stacking cores can improve the die yield, there are losses on account of joining 2 layers together, quantified by the *bonding yield*, as shown in equation 24 in [5]. As a result there is a tradeoff between increasing the die size and increasing the number of layers.

In order to counteract this yield variation for a multilayered processor system, we consider the use of redundant or spare cores. For a given multicore system, we consider only a subset of the total cores on chip to be operational. The remaining cores are used to ensure that a minimum yield requirement of 50% is met. This is known as *core sparing* and is commonly used as a technique to improve the overall yield of several processors in industry [11]. Although additional hardware resources are spent on these spare cores, the improvement in yield significantly shortens the time to market for these processors. This is a far more viable alternative than aiming to improve the fabrication process both from a time and cost perspective. Adding a single spare core to an 8 core system can reduce the time to market by nearly 50%. However the number of spare cores needed to meet the yield criteria increases with the total number of cores. For larger number of cores and layers, the yield drops drastically, resulting in more than 50% of the cores being used for redundancy.

Figure 6 shows the number of stacked layers required to obtain a particular number of cores for different area footprints. The redundancy ratio is defined as the fraction of excess cores required to meet the yield threshold. It can be observed that both area footprint and number of layers cause this redundancy ratio to increase. For smaller areas the reduction in yield due to bonding is a more dominant characteristic, as indicated by the increase in redundancy ratio for more stacked layers. However, as the area per layer increases, the yield decreases at a faster rate and folding the cores to stack them in multiple layers can arrest this decline. The maximum area footprint considered is 400 mm^2 per layer. In order to meet the yield constraint, it is essential to increase the number of layers to accommodate the redundant cores. This adversely affects the thermal behavior of the processor, further constraining the design space.

An important advantage that TFET technology has over other emerging devices is that it is compatible with the CMOS fabrication process [9]. Further, the process steps involved in the manufacture of TFET processors is similar to that of CMOS. As a result, we assume that this technology is similarly affected by process variation and displays similar yield as CMOS [32]. In order to account for uncertainties due to the new technology, we ran yield experiments for TFET processors by reducing the baseline yield by 5% and 10%. The reduction in overall yield could be compensated by adding an additional redundancy of 8% and 17% respectively, without compromising the feasible design space for TFET processors.

4.3. Modeling thermal distribution across multicores

In addition to affecting processor reliability and lifetime, the cost of cooling and packaging is determined by the the thermal profile of the processor. Different cooling technologies such as microfluidic cooling can push this thermal limit up. For instance microfluidic cooling techniques for 3D processor-on-processor stacking can reduce core temperature by as much as 15°C [42]. For the purpose of this study we consider a thermal limit of around 85-90°C (358-363K), assuming an air cooled machine. Microfluidic cooling could enable the temperature bound to be raised to around 100-105°C.

4.4. Variation in microarchitecture

As part of our studies we evaluated the effect of microarchitecture changes under the thermal constraint. A simpler and narrower-issue processor will consume less power than a wider, more complex core. As a result, there is more thermal

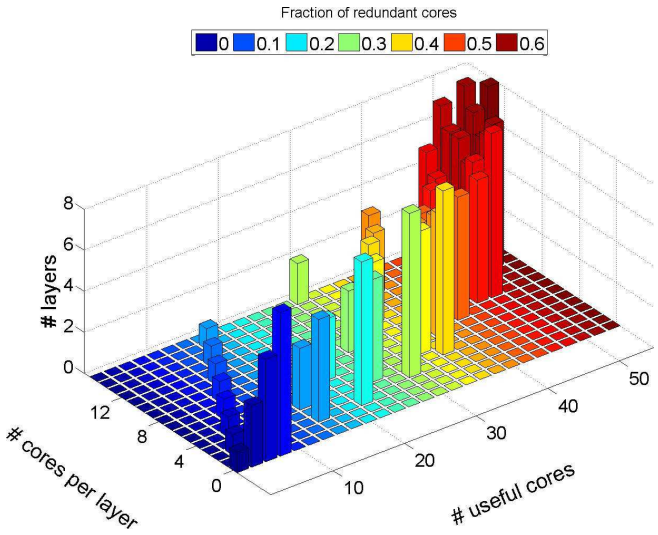

Figure 6: Number of core layers required to realize a range of functioning cores for different area footprints. The fraction of redundant cores can be seen to increase both with area and with number of layers

slack available, which could enable a higher frequency operation than the complex processor. Thus it could be possible to match the performance of the complex core using a simpler core configuration, under similar thermal constraints. However, preliminary experiments (not shown) indicated that the performance loss from moving from out-of-order to in-order execution outweighs power and thermal advantages for the space we consider. Thus, we employ a low-issue out-of-order configuration as our simpler core. Our experiments were carried out on a 2-issue Atom-like core configuration and a 4-issue Ivybridge-like microarchitecture.

The ability to exploit the greater microarchitecture complexity or increased number of cores depend on the application characteristics, in particular the ILP and TLP of the application. As discussed in Section 2, there is a region in the parallelism v/s frequency plot that is not attainable because of yield and thermal constraints. By using a combination of power-efficient TFET technology with high performance CMOS can expand the design space. Using TFET cores in conjunction with 3D technology can reduce the power consumed, while maintaining processor yield, thus mitigating the thermal constraint. However, whether this extra design space manifests itself as a performance improvement depends entirely on the application scaling behavior.

Figure 7a) and b) show the 2 applications, *barnes* and *ocean.cont*, respectively, that represent the extreme edges of application scaling with cores. The additional TFET cores operating at low frequency, would prove beneficial for highly parallel applications like *barnes* which is then able to improve its peak performance, as seen in Figure 7a) . On the other hand, an application like *ocean.ncont*, shown in Figure 7b), which has limited TLP, prefers operating on fewer cores at higher frequency. It is evident that *barnes* benefits greatly from the extra number of cores, whereas the effect is not very significant in *ocean.ncont*.

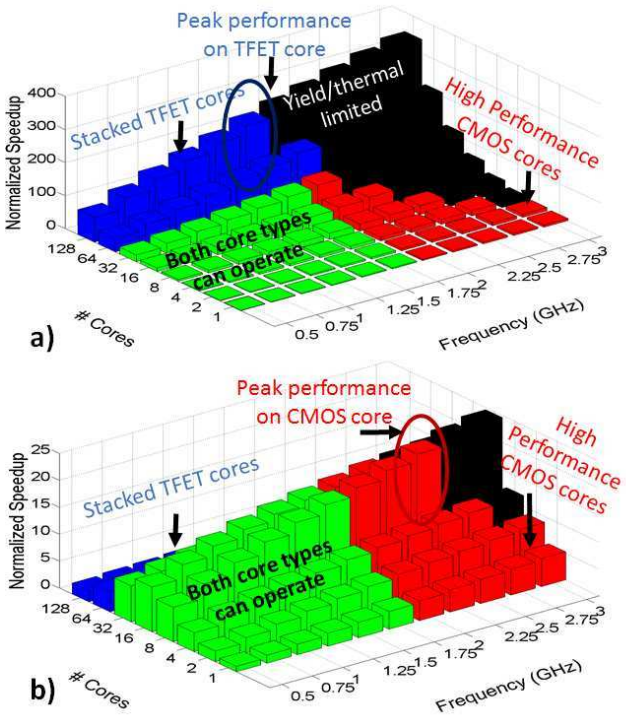

Figure 7: a) and b) Delineation of design space attainable by CMOS(red), TFET(blue), both (green) and neither (black) cores to obtain peak performance, for a scalable (*barnes*) and non-scalable (*ocean.ncont*) application respectively. The best performance is seen in the TFET configuration in *barnes* and in the CMOS configuration in *ocean.ncont*.

5. System Infrastructure and Simulation Tools

Our simulations were carried out on a multicore system comprising primarily of Intel Ivybridge-like cores. Table 2 lists the system configuration that we ran for our default simulations

Processor	Ivybridge Microarchitecture
L1 Cache	32 KB D/I, 8 way S.A
L2 Cache	256 KB private Cache
L3 Cache	2MB/Core shared LLC, 16 way S.A
DRAM	4GB, DDR3-1600, 1 mem channel

Table 2: Configuration of the evaluation platform.

5.1. Architectural simulation setup and benchmarks

We used the Sniper-5.0 [4] system simulation tool for our performance simulations. This tool is integrated with McPAT-0.8 [24] which is used for power and area estimation. We instrumented the McPAT technology file with parameters obtained from our TCAD simulations described in Section 3. These parameters are listed in Table 3. In addition McPAT

interfaces with Cacti, providing timing models for every processor component as well as wires, which we use to determine the critical path delay.

Parameter	FinFET	HTFET	Parameter	FinFET	HTFET
C_{ox}(fF/μm)	1.28	1.28	L_g(nm)	26	26
V_{th}(V)	0.25-0.3	0.1 (eff)	V_{d-sat}(V)	0.419	0.288
R_{on}(Kω-μm)	1.01	2.43	I_{on}(mA-μm)	0.71	0.166
$C_{g-ideal}$(fF/μm)	0.55	0.327	EOT(nm)	0.7	0.7
Source Doping($/cm^3$)	1e20 n+	4e19 GaSb p+	Drain Doping($/cm^3$)	1e20 n+	8e17 InAs n+

Table 3: Technology parameters.

For the purpose of obtaining thermal profiles, we created periodic traces using Sniper and created a power profile by running McPAT on each individual trace. Our processor logic and wire models, described in Section 4 were used to obtain the corresponding TFET numbers from the CMOS core simulations. These power profiles were then used as input to Hotspot3D for obtaining temperature variations across the processor.

5.2. Modeling thermal variation

For determining the thermal distribution across a 3D stacked multicore, we used the *Hotspot-5.02* tool [16]. As shown in [3], the power budget of a multicore is based on its thermal profile, which in turn, depends on the temperature (or power) distribution across adjacent cores. This is done in order to take into account the effect of heat dissipation across core boundaries. Hence, in order to model this phenomenon, we carried out simulations on a system of 3 cores arranged side-by-side, each one operating on a similar workload. The central core will naturally experience the highest temperature distribution, since the surrounding cores operating at high temperature offer limited avenues for heat dissipation. Secondary effects beyond multiple core boundaries are negligible and can hence be ignored. Figure 8 demonstrates the variation in thermal behavior of three such 4-issue CMOS cores running at 2 GHz. The central cores in each layer have a larger area of peak temperature due to its higher power density as compared to the cores at the boundary. This model can be replicated to correspond to several cores on the same layer. We then consider a 4 layered multicore with a (20 μm) thin thermal insulating material between each core layer. We then assume that this model is extended to 3D, with each layer comprising of a multitude of such cores. The additional temperature increase due to transition to 3D is modeled in Hotspot3D [28].

5.3. Scheduling diverse workloads on a stacked CMOS-TFET multicore

We aim to demonstrate the motivation behind using a heterogeneous 3D stacked multicore for efficiently executing a diversity of applications. This configuration is comprised of a single (top) layer consisting of CMOS cores and remaining layers consisting of TFET cores. We use a selection of multiprogrammed workloads created from the *Parsec*, *Splash2* and *SPEC CPU2006* suites for this purpose. In order to carry

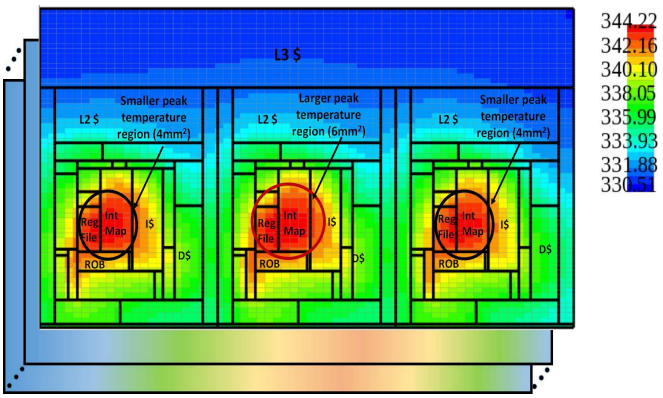

Figure 8: Variation in thermal hotspots on a layer of a system of 3D stacked cores with shared L3 cache.

out a comprehensive study of the different workload characteristics encountered by the system, we statically profile each benchmark and determine their thread-level parallelism and memory utilization. We then partition them into various subgroups depending on these characteristics. These benchmarks are characterized as the following:

1. Multithreaded - High scalability
2. Multithreaded - Limited scalability
3. 1-threaded (no scalability) with high memory utilization
4. 1-threaded (no scalability) with low memory utilization

Figure 9 shows the benchmarks that we profiled and used for obtaining representative workloads.

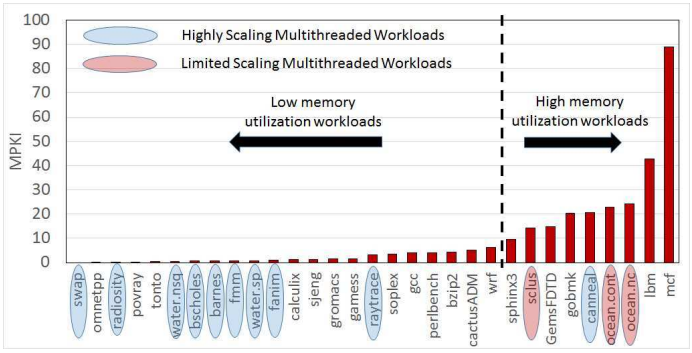

Figure 9: Characterization of *SPEC CPU2006*, *Parsec* and *Splash2* based on scalability and memory utilization

By randomly combining pairs of workloads from these categories, we get several distinct classes of multiprogrammed workloads. The thermal constraint under which the multicore configuration operates primarily holds when majority of cores are active. On the other hand, the shared L3 cache, having a much smaller activity factor does not consume as much power as cores and is consequently much cooler. This enables us to operate the on-chip caches when majority of cores are turned off, during execution of single threaded applications. Thus, depending on the workload characteristics, we can either operate the system as a 3D stacked processor multicore or as

a single (or multiple) layer of cores with a large stacked L3 cache.

Figure 10 illustrates the heterogeneous configuration that we propose and the possible states it can operate under for different workloads. The assignment of cores is as shown in the figure. Depending on the memory utilization as determined previously, the cache is partitioned according to the following heuristic, which we term *Heterogeneity Aware Scheduler (HAS)*.

- The L3 cache local to the core is initially allocated to the application running on that core.
- The remaining L3 cache local to unused cores are preferentially allocated to applications depending on whether they are sensitive to cache size or not, as determined by their *MPKI* (Misses per Kilo Instruction).
- If both applications have high cache utilization, then the un-allocated L3 cache is partitioned equally between applications.
- If both applications have poor cache utilization, then the un-allocated L3 cache is left unused, in order to preserve locality and reduce access latency.
- If the applications have differing L3 cache utilization characteristics, the application with higher utilization is allocated the unused cache.

We only carry out a binary classification based on cache utilization as either dependent or independent of cache size. and avoid finer grained comparisons of relative utilization. This is because the true working set size of an application can be highly data dependent and the response of the application to increasing or decreasing the cache size may not be deterministic.

Thus, we propose the following algorithm that optimally utilizes the on-chip resources for a wide variety of applications to maximize the thermally constrained performance. Each application is profiled *a priori* to determine its scalability and memory utilization.

We used a combination of workloads from the *Splash2*, *Parsec* [2] and *SPEC CPU2006* suites. For our heterogeneous scheduling experiments, we used random combinations of workloads with different scaling and memory utilization characteristics, as described in Section 4. Table 4 shows the workload mixes that we evaluated.

Workload-mix	W1 characteristic	W2 characteristic
	Scaling, MPKI	Scaling, MPKI
mcf-gobmk	No , high	No , high
lbm-scluster	No , high	Weakly , high
mcf-canneal	No , high	Strongly , high
gcc-sphinx	No , low	No , high
barnes-fanim	Strongly , low	Strongly , low
ocean.nc-raytrace	Weakly , high	Strongly , low
ocean.c-scluster	Weakly , high	Weakly , high
canneal-ocean.ncont	Strongly , high	Weakly , high

Table 4: Configuration of the evaluation platform.

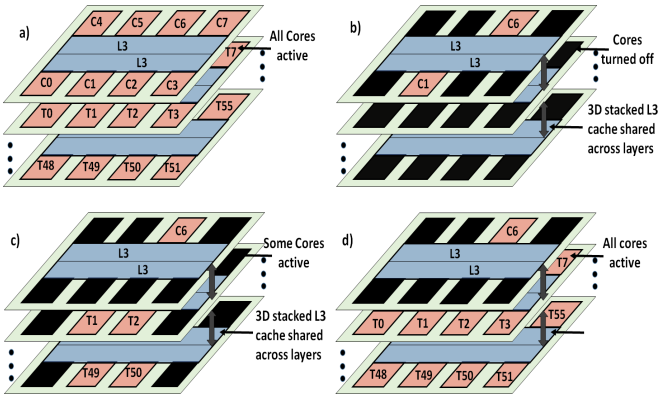

Figure 10: Different operating states of the heterogeneous multicore: a) 2 highly scalable parallel applications scheduled on the entire multicore. b) 2 completely sequential applications scheduled exclusively on CMOS cores. c) A sequential application, scheduled on a CMOS core, running alongside a weakly scaling application. The latter is scheduled on either the remaining CMOS cores or TFET cores depending its optimal configuration. d) A sequential application, scheduled on a CMOS core, running alongside a highly parallel application scheduled on the entire set of TFET cores.

6. Results

In this section, we carry out a design space exploration across different architecture, device and system configurations for a diversity of workloads. We attempt to find the best possible device-architecture co-design for each of these workloads. Our studies are carried out by varying the core frequency and the number of cores across multiple stacked layers, under the thermal and yield constraints described in Section 4. We also demonstrate sensitivity results over a range of temperature budgets and microarchitectural configurations.

6.1. Optimal operating points in the design space

Figure 11a) and b) show the various optimal design points that are possible for different sets of applications (*parsec* and *splash2* respectively). The harmonic mean of the relative speedups of all applications in each benchmark suite (relative to a single core, operating at peak frequency), at every operating point is evaluated for each category. In addition to the red and blue bars, which signify the operating points exclusive to CMOS and TFET cores respectively, the green colored bars denote all states that can be attained by both core types. The diversity in the overall scalability of the workload suite is evident in the comparison between Figures 11a) and 11 b). In order to determine which core configuration is preferred in the green region, we determine the CMOS and TFET power for all states in this region and plot the power savings obtained by using one core over the other, as shown in Figure 12. In this figure, the red and blue regions correspond to those core states where it is more power efficient to use CMOS or TFET respectively. This plot clearly illustrates that for optimal performing designs in the TFET-preferred region, the power sav-

ings can be significant and as shown in Figure 5, the relative gains w.r.t CMOS will increase with subsequent generations.

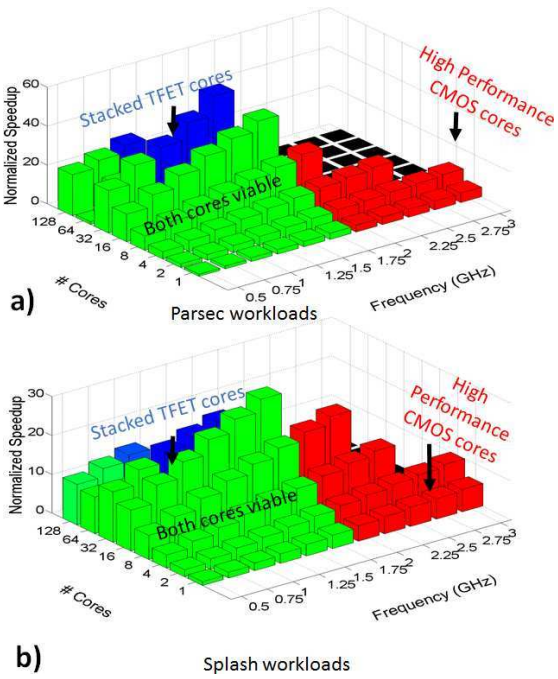

Figure 11: a) and b) Mean speedup of different applications in the *splash2* and *parsec* suites respectively. The *splash2* applications, on average, prefer higher frequency and fewer cores to operate (red CMOS space and green common space). This is because ∼29% of applications are relatively poor scaling. On the other hand, *parsec* benchmarks operate most efficiently on larger number of cores and lower frequencies (blue TFET space and green common space), with only ∼17% applications preferring high frequency CMOS cores

Figure 13 shows the performance comparison of a 4-issue Ivybridge TFET v/s CMOS processor for a range of *Splash2* and *Parsec* benchmarks and compares the best performing configurations in each case. The optimal configurations for each processor (frequency, number of layers), subject to thermal and yield constraints are indicated for each data point. All speedups are normalized to a single CMOS core running at peak frequency (3 GHz). The TFET core configurations outperform the best CMOS configuration by an average of around 17% for the *Splash2* suite and around 20% for the *Parsec* suite. The overall speedup is around 18%. This performance improvement varies with the temperature budget as shown below.

Table 5 shows the best performing configuration under thermal constraints in terms of frequency, number of cores and number of stacked layers for both CMOS and TFET processors.

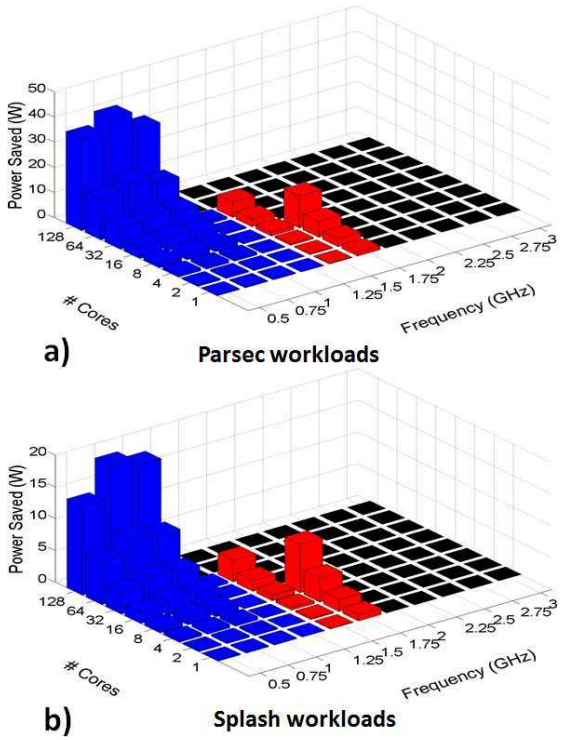

Figure 12: a) and b) Mean power savings obtained in *parsec* and *splash2* suites respectively by running workloads on CMOS (red region) or TFET (blue region)cores. This plot is restricted only to the region where both CMOS and TFET cores are capable of operation in order to determine the more efficient core.

Benchmark	Technology	Frequency(GHz)	Cores	Layers
SPLASH				
barnes	CMOS	1	64	8
	TFET	1.25	64	8
fmm	CMOS	1	64	8
ocean.cont	CMOS	1.75	32	4
	TFET	1.5	32	4
ocean.ncont	CMOS	1.75	32	4
	TFET	1.5	32	4
radiosity	CMOS	1	64	8
	TFET	1.25	64	8
	TFET	1.25	64	8
water.nsq	CMOS	1.5	32	4
	TFET	1.25	64	8
water.sp	CMOS	1	64	8
	TFET	1.25	64	8
PARSEC				
blackscholes	CMOS	1	64	8
	TFET	1.25	64	8
canneal	CMOS	1	64	8
	TFET	1.25	64	8
fanimate	CMOS	1	64	8
	TFET	1.25	64	8
raytrace	CMOS	1	64	8
	TFET	1.25	64	8
scluster	CMOS	1.75	32	4
	TFET	1.5	32	4
swaptions	CMOS	1	64	8
	TFET	1.25	64	8

Table 5: Best performing configuration for each workload

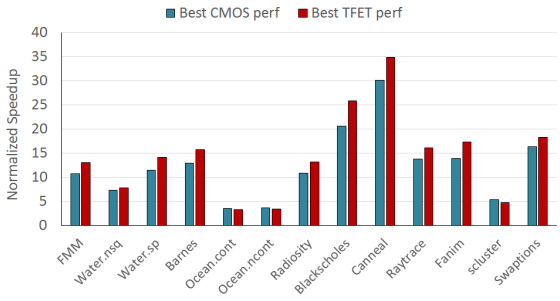

Figure 13: Relative performances of 3D stacked CMOS and TFET configurations using an 8 stacked layers consisting of 64 functioning 4 issue processors. The thermal budget assumed here is 87°C.

6.2. Sensitivity to thermal budget

Figure 14 shows the variation in performance improvement obtained by TFET by comparing the best possible TFET and CMOS core configurations. TFET cores are clearly the preferred choice for thermal budgets upto around 360K (87°C), while the performance difference is negligible upto around 380K (107°C). At higher thermal budgets above a 100°C, CMOS cores clearly dominate since, the scope of microarchitectural configurations that they can attain is large enough to offset the increased thermal efficiency of TFET cores.

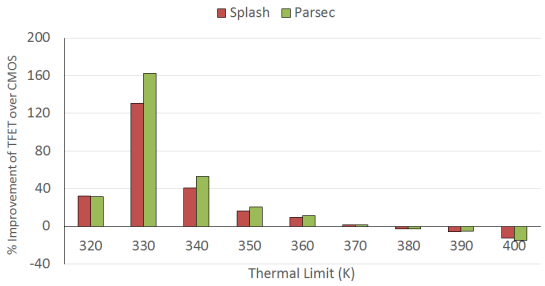

Figure 14: Variation of performance improvement of TFET core as opposed to CMOS cores for different thermal limits. Evaluations are carried out separately for *Splash* and *PARSEC* benchmark suites

6.3. Sensitivity to microarchitecture

We also carried out experiments that compare CMOS and TFET performance for a range of processor microarchitectures, ranging from a single issue to an 8-issue out-of-order processor, as shown in Figure 15. This figure shows the mean speedup of all benchmarks run from the *Splash2* and *Parsec* suites, when compared to a single core baseline running at peak frequency (3 GHz). The 3D stacked multicore configuration remains the same as the previous experiments and is subjected to the same thermal limit of 360K. The *Splash* suite of benchmarks are not very sensitive to processor complexity as there is only a minor improvement in speedup with increase in issue width. On the other hand, in case of the *Parsec* suite, the performance improvement of TFET processors peaks at the 4 issue configuration. This is because the 4-issue TFET processor has sufficient capacity to exploit the inherent ILP of the application. As a result, when combined with 3D stacking, this configuration is able to extract the maximum performance from the application by optimizing both its ILP and TLP. For lower issue processors, core frequency plays a more important role, which reduces the advantage due to TFET cores. On the other hand, wider (6 and 8 issue) processors are extremely power hungry and provide limited improvement in ILP over the 4 issue configurations. As a result, the higher base temperature attained by these cores, severely limits the microarchitural flexibility in both CMOS and TFET cores, leading to lower speedups.

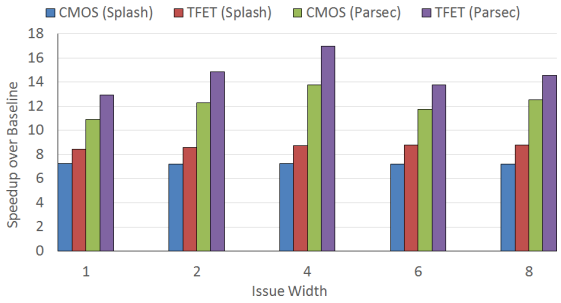

Figure 15: Comparison of performance speedup of CMOS and TFET cores for different microarchitectural configurations. Evaluations are carried out separately for *Splash2* and *Parsec* benchmark suites for issue widths of 1 to 8

6.4. Heterogeneity aware scheduling

In addition to the static profiling based results, we also demonstrated the viability of a stacked CMOS-TFET heterogeneous multicore. We implemented our *Heterogeneity-Aware Scheduler*, by running the workload mixes in 4 on the CMOS-TFET multicore and compared the improvement in performance to the both a homogeneous CMOS and a homogeneous TFET multicore. The results are shown in Figure 16. All results are weighted speedups normalized to the ideal baseline, i.e the weighted speedup of each application when run individually on the best possible CMOS/TFET configuration. The heterogeneity aware scheduler results in a 17% improvement over the best homogeneous configuration.

7. Related Work

7.1. Architectures for processors in emerging technologies

Heterogeneous integration of CMOS with several emerging devices has been an active research area due to the shortcomings being exposed in conventional CMOS with technology scaling. [17] has proposed the use of emerging technologies like Nano-Electro-Mechanical FETs and Tunnel FETs in integrated circuit logic. In this paper, we extend device level models to a complete processor-level abstraction, including logic and wire delay and power modeling. The use of

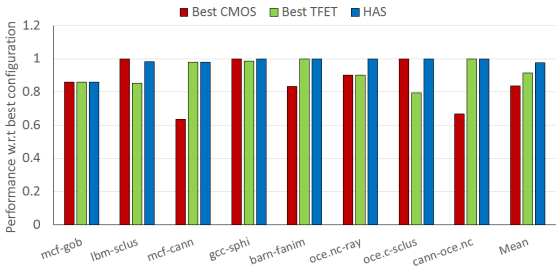

Figure 16: Performance comparison of homogeneous CMOS and TFET multicore with a heterogeneous 3D configuration with HAS, consisting of 1 CMOS layer and 7 TFET layers

device-level heterogeneous multicores comprising of CMOS and TFET based processors has been explored in [36], [21] and [37]. These papers mainly propose the use of simple, architecturally homogeneous multicores, with application mapping so as to minimize power under an iso-performance constraint or maximize performance in an iso-power scenario. This paper explores design of 3D stacked processors comprising TFET and CMOS cores and explore different architectural tradeoff points under thermal and yield constraints.

7.2. 3D stacked architectures

The authors in [13] analyze various aspects of heterogeneous device integration in 3D technology. Our work, while based on this concept, extends this further by incorporating thermal and yield models as well as application mapping algorithms on heterogeneous 3D architectures. [15] proposes the idea of *Resource Pooling* across multiple stacked layers within a processor and efficiently allocating resources and tasks to each layer. Our work examines the viability of stacking processors on top of each other without folding individual cores which is an extremely complicated task due to wiring concerns.

7.3. Heterogeneous architectures

Heterogeneous Asymmetric Chip Multi Processors cores have been proposed in the past [22, 23] which deal with the varying demands from the applications in terms of Instruction Level Parallelism (ILP) and Thread Level Parallelism (TLP). These schemes decide the number of big (out-of-order) and small (in-order) cores statically. To circumvent this static techniques, there have been works [20], [18] which dynamically transform the issue widths to cater to the sequential (ILP) and parallel(TLP) parts of the applications. [18] employs dynamically fusing the cores to form a large one core group to increase the performance of the sequential code. These fusable cores can also execute parallel threads independently in isolation to speed up the parallel portions of the application. This technique suffers from the overheads while reconfiguring like instruction cache flushes, data migration etc. which are circumvented in [20] by dynamically transforming an Out of order core in to SMT based in-order core.

7.4. Architectural techniques for power and thermal aware execution

Dreslinski, *et. al* [10] proposed the concept of *Near Threshold Computing* (NTC), where the processors are run at a voltage level close to V_t, so that it operates at a point where the power efficiency is the highest. By incorporating TFET cores into our system, the near-threshold voltage restriction is eliminated and a far wider range of operating points are available to the user. *Computational Sprinting* is another technique which can be used to improve the utilization of processors under power and thermal limitations [30]. In this paper, the authors briefly allow the processor to briefly exceed the processor power limitation by operating it at extremely high performance points for short periods of time. Our system allows for high performance CMOS cores to selectively run based on the application requirement, which covers most of the scenarios for which sprinting is required. It is also a more robust technique from the perspective of reliability and aging.

7.5. Thermal-Aware application mapping on multicores

In [35], the authors propose PROMETHEUS, a heterogeneous multiprocessor SOC-based thermal-aware scheduling policies, such as TempoMP and TemPrompt. Our techniques extend over a wide range of application domains and incorporated everywhere from embedded SOCs to high end server architectures. [14] describes a thermal-aware DVFS algorithm for real-time applications running on a multicore, while [6] proposes DVFS techniques on 3D multicores. This paper examines general purpose applications with a large diversity in scaling behavior and memory utilization and attempts to optimize the thermally constrained performance on a device-level heterogeneous multicore, for the entire spectrum of application characteristics.

8. Conclusion

In this paper, we conducted an extensive evaluation of the performance and energy tradeoffs among choices in device technologies, 3D integration, microarchitecture, and scheduling under constraints imposed by realistic yield models, thermal bounds and exploitable application parallelism. We showed that, with 3D integration, steep-slope based devices can extend the space of viable designs and achieve competitive performance for highly parallel applications. We highlight how, with further technology scaling, the range of applications for which steep-slope devices are appropriate grows, while the portion of the design space where CMOS is optimal shrinks to a point where only a small number of CMOS cores may be desirable. Finally, we present a method for capitalizing on these trends by developing an intelligent scheduling approach for hybrid CMOS-TFET systems that yields mean improvements ranging from 17% for a high end server workloads running at around 90°C to over 160% for embedded systems running below 60°C.

Acknowledgments

This work was supported by the Center for Low Energy Systems Technology (LEAST), one of six centers supported by the STARnet phase of the Focus Center Research Program (FCRP), an SRC program sponsored by MARCO and DARPA. It was also supported by the NSF ASSIST NERC award 1160483 and NSF awards 1205618 and 1213052. The authors would also like to thank Prof. Suman Datta and his group at Penn State University for his collaboration on TFET devices and models used in this work.

References

[1] "TCAD Sentaurus Device Manual," 2010.

[2] C. Bienia and K. Li, "PARSEC 2.0: A new benchmark suite for chip-multiprocessors," in *Proceedings of the 5th Annual Workshop on Modeling, Benchmarking and Simulation*, 2009.

[3] A. Branover, D. Foley, and M. Steinman, "AMD Fusion APU: Llano," *Micro, IEEE*, vol. 32, no. 2, pp. 28–37, 2012.

[4] T. Carlson, W. Heirman, and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation," in *High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference for*, 2011, pp. 1–12.

[5] Y. Chen, D. Niu, Y. Xie, and K. Chakrabarty, "Cost-effective integration of three-dimensional (3D) ICs emphasizing testing cost analysis," in *Computer-Aided Design (ICCAD), 2010 IEEE/ACM International Conference on*, 2010, pp. 471–476.

[6] H. J. Choi, Y. J. Park, H.-H. Lee, and C. H. Kim, "Adaptive dynamic frequency scaling for thermal-aware 3D multi-core processors," in *Proceedings of the 12th International Conference on Computational Science and Its Applications - Volume Part IV*, ser. ICCSA'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 602–612. Available: http://dx.doi.org/10.1007/978-3-642-31128-4_44

[7] L. Czornomaz *et al*., "Co-integration of InGaAs n- and SiGe p-MOSFETs into digital CMOS circuits using hybrid dual-channel ETXOI substrates," in *Electron Devices Meeting (IEDM), 2013 IEEE International*, Dec 2013, pp. 2.8.1–2.8.4.

[8] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *Solid-State Circuits, IEEE Journal of*, vol. 9, no. 5, pp. 256–268, 1974.

[9] G. Dewey *et al*., "Fabrication, characterization, and physics of III-V heterojunction tunneling field effect transistors (H-TFET) for steep sub-threshold swing," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, 2011, pp. 33.6.1–33.6.4.

[10] R. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits," *Proceedings of the IEEE*, 2010.

[11] P. Emma *et al*., "3D stacking of high performance processors," in *High Performance Computer Architecture (HPCA) Industry Session, 2014 IEEE 20th International Symposium on*, 2014, pp. 1–12.

[12] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark Silicon and the End of Multicore Scaling," in *Proceedings of the 38th International Symposium on Computer Architecture (ISCA)*, 2011.

[13] M. Fernandez-Bolanos and A. Ionescu, "3D heterogeneous integration for novel functionality," in *3D Systems Integration Conference (3DIC), 2010 IEEE International*, 2010, pp. 1–19.

[14] V. Hanumaiah and S. Vrudhula, "Temperature-aware DVFS for hard real-time applications on multicore processors," *Computers, IEEE Transactions on*, vol. 61, no. 10, pp. 1484–1494, 2012.

[15] H. Homayoun, V. Kontorinis, A. Shayan, T.-W. Lin, and D. Tullsen, "Dynamically heterogeneous cores through 3D resource pooling," in *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*, 2012, pp. 1–12.

[16] W. Huang *et al*., "Accurate pre-RTL temperature-aware design using a parameterized, geometric thermal model," in *ISSCC*, 2008.

[17] A.-M. Ionescu *et al*., "Ultra low power: Emerging devices and their benefits for integrated circuits," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, 2011, pp. 16.1.1–16.1.4.

[18] E. Ipek, M. Kirman, N. Kirman, and J. F. Martinez, "Core fusion: accommodating software diversity in chip multiprocessors," in *Proceedings of the 34th annual international symposium on Computer architecture*, 2007.

[19] ITRS, "The international technology roadmap for semiconductors (ITRS)," 2011.

[20] Khubaib, M. A. Suleman, M. Hashemi, C. Wilkerson, and Y. N. Patt, "Morphcore: An energy-efficient microarchitecture for high performance ILP and high throughput TLP," in *MICRO*, 2012.

[21] E. Kultursay, K. Swaminathan, V. Saripalli, V. Narayanan, M. Kandemir, and S. Datta, "Performance enhancement under power constraints using heterogeneous CMOS-TFET multicores," in *CODES*, 2012.

[22] R. Kumar, K. Farkas, N. Jouppi, and P. Ranganathan, "Single-isa heterogeneous multi-core architectures: The potential for processor power reduction," in *MICRO*, ser. MICRO 36, 2003.

[23] R. Kumar, D. Tullsen, P. Ranganathan, N. Jouppi, and K. Farkas, "Single-ISA heterogeneous multi-core architectures for multithreaded workload performance," in *ISCA*, 2004.

[24] S. Li *et al*., "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *MICRO*, 2009.

[25] L. Liu *et al*., "Scaling length theory of double-gate interband tunnel field-effect transistors," *Electron Devices, IEEE Trans*., 2012.

[26] L. Liu, D. Mohata, and S. Datta, "Scaling length theory of double-gate interband tunnel field-effect transistors," *Electron Devices, IEEE Transactions on*, vol. 59, no. 4, pp. 902–908, 2012.

[27] Z. Lu *et al*., "Realizing super-steep subthreshold slope with conventional FDSOI CMOS at low-bias voltages," in *IEDM*, Dec.

[28] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked dram under power and thermal constraints," in *Proceedings of the 49th Annual Design Automation Conference*, ser. DAC '12. New York, NY, USA: ACM, 2012, pp. 648–655. Available: http://doi.acm.org/10.1145/2228360.2228477

[29] S. Mookerjea *et al*., "Experimental Demonstration of 100nm Channel Length In0.53Ga0.47As-based Vertical Inter-band Tunnel Field Effect Transistors (TFETs) for Ultra Low-Power Logic and SRAM Applications," in *IEDM*, 2009.

[30] A. Raghavan *et al*., "Utilizing dark silicon to save energy with computational sprinting," *Micro, IEEE*, vol. 33, no. 5, pp. 20–28, 2013.

[31] R. Rooyackers *et al*., "A new complementary hetero-junction vertical tunnel-FET integration scheme," in *Electron Devices Meeting (IEDM), 2013 IEEE International*, Dec 2013, pp. 4.2.1–4.2.4.

[32] V. Saripalli, S. Datta, V. Narayanan, and J. Kulkarni, "Variation-tolerant ultra low-power heterojunction Tunnel-FET SRAM design," in *Nanoscale Architectures (NANOARCH), 2011 IEEE/ACM International Symposium on*, 2011, pp. 45–52.

[33] V. Saripalli *et al*., "An Energy-Efficient Heterogeneous CMP based on Hybrid TFET-CMOS cores," in *DAC*, 2011.

[34] A. C. Seabaugh and Q. Zhang, "Low-voltage tunnel transistors for beyond CMOS logic," *Proceedings of the IEEE*, Dec.

[35] S. Sharifi *et al*., "Prometheus: A proactive method for thermal management of heterogeneous MPSoCs," *TCAD*, 2013.

[36] K. Swaminathan, E. Kultursay, V. Saripalli, V. Narayanan, M. Kandemir, and S. Datta, "Improving energy efficiency of multi-threaded applications using heterogeneous CMOS-TFET multicores," in *ISLPED*, 2011.

[37] K. Swaminathan, E. Kultursay, V. Saripalli, V. Narayanan, M. Kandemir, and S. Datta, "Steep-slope devices: From dark to dim silicon," *Micro, IEEE*, vol. 33, no. 5, pp. 50–59, 2013.

[38] M. B. Taylor, "Is dark silicon useful?: Harnessing the four horsemen of the coming dark silicon'apocalypse," in *DAC*, 2012.

[39] K. Tomioka, M. Yoshimura, E. Nakai, F. Ishizaka, and T. Fukui, "Integration of III-V nanowires on Si: From high-performance vertical FET to steep-slope switch," in *Electron Devices Meeting (IEDM), 2013 IEEE International*, Dec 2013, pp. 4.1.1–4.1.4.

[40] G. Venkatesh *et al*., "Conservation cores: reducing the energy of mature computations," in *ASPLOS*, 2010.

[41] X. Wu *et al*., "Cost-driven 3D integration with interconnect layers," in *Proceedings of the 47th Design Automation Conference*, ser. DAC '10. New York, NY, USA: ACM, 2010, pp. 150–155. Available: http://doi.acm.org/10.1145/1837274.1837313

[42] Y. Zhang, A. Dembla, Y. Joshi, and M. Bakir, "3D stacked microfluidic cooling for high-performance 3D ICs," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, 2012, pp. 1644–1650.